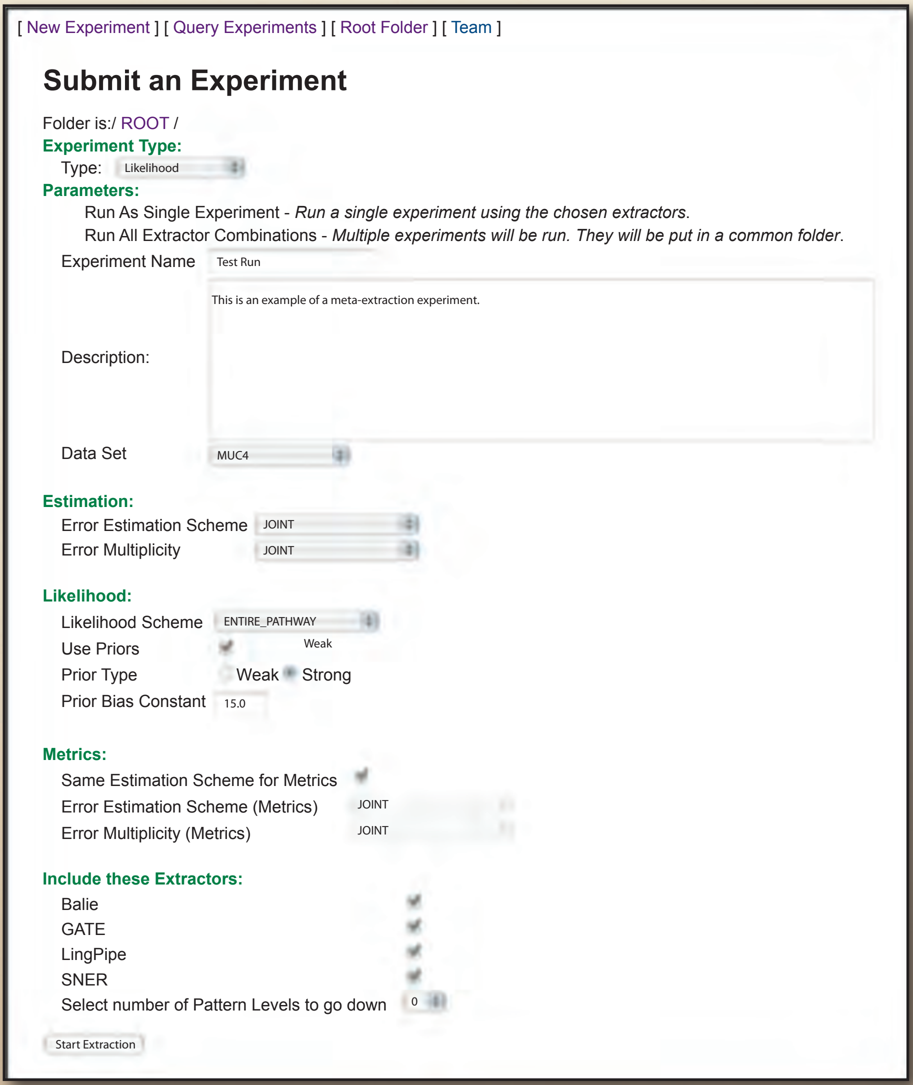
Entity Extractor Aggregation System

Tracy D. Lemmond
(925) 422-0219
lemmond1@llnl.gov

The extraction of relational information (such as triples and events) and entities (such as people and organizations) from unstructured text often forms the basis for data ingestion by Knowledge Discovery (KD) systems. These systems enable analysis and inference on massive sets of data and are particularly vulnerable to errors introduced during the ingestion process.



[New Experiment] [Query Experiments] [Root Folder] [Team]

Submit an Experiment

Folder is:/ ROOT /

Experiment Type:

Type: Likelihood

Parameters:

Run As Single Experiment - *Run a single experiment using the chosen extractors.*
Run All Extractor Combinations - *Multiple experiments will be run. They will be put in a common folder.*

Experiment Name: Test Run

Description: This is an example of a meta-extraction experiment.

Data Set: MUC4

Estimation:

Error Estimation Scheme: JOINT
Error Multiplicity: JOINT

Likelihood:

Likelihood Scheme: ENTIRE_PATHWAY
Use Priors ☑ Weak
Prior Type ○ Weak ● Strong
Prior Bias Constant: 15.0

Metrics:

Same Estimation Scheme for Metrics ☑
Error Estimation Scheme (Metrics): JOINT
Error Multiplicity (Metrics): JOINT

Include these Extractors:

Balie ☑
GATE ☑
LingPipe ☑
SNER ☑
Select number of Pattern Levels to go down: 0

Figure 1. Parameter selection for a meta-extractor experiment.

Though state-of-the-art extraction tools achieve insufficient accuracy rates for practical use, not all extractors are prone to the same types of error. This suggests that improvements may be achieved via appropriate combinations of existing extraction tools, provided their behavior can be accurately characterized and quantified. Several methodologies that combine pattern-based and probabilistic approaches exist to address the entity extraction problem via the aggregation of extraction tools (*i.e.*, base extractors). The focus of this effort has been to construct an operational prototype of these algorithms within an extensible framework that will allow future advancements to be easily incorporated and enable the generation of detailed results and performance assessment data.

Project Goals

Key objectives for this system include 1) building a "plug-and-play" extensible framework to enable the incorporation of future algorithmic advancements; 2) automating the execution of experiments and the navigation of experimental results; 3) providing tools for evaluating the performance of base extractors and of the final aggregated output; and 4) incorporating methodology characterization tools that facilitate algorithm optimization.

Relevance to LLNL Mission

Nonproliferation, counterterrorism, and other national security missions rely on the acquisition of knowledge that is buried in unstructured text documents too numerous to be manually processed. Systems are being worked on by LLNL and its customers that must automatically extract entities from these sources. Methodologies have been created that significantly advance entity extraction capabilities. Bringing these capabilities to an operational status is critical to the timely deployment of KD technologies that will impact LLNL mission goals. This effort directly supports LLNL's Engineering Systems for Knowledge and Inference (ESKI) Text to Inference area and the Cyber, Space, and Intelligence strategic mission thrust in the LLNL five-year strategic roadmap. The completed system will provide highly valued and unprecedented entity extraction capabilities to internal programs, such as IOAP and CAPS, and to external customers such as DHS, DoD, and the IC.

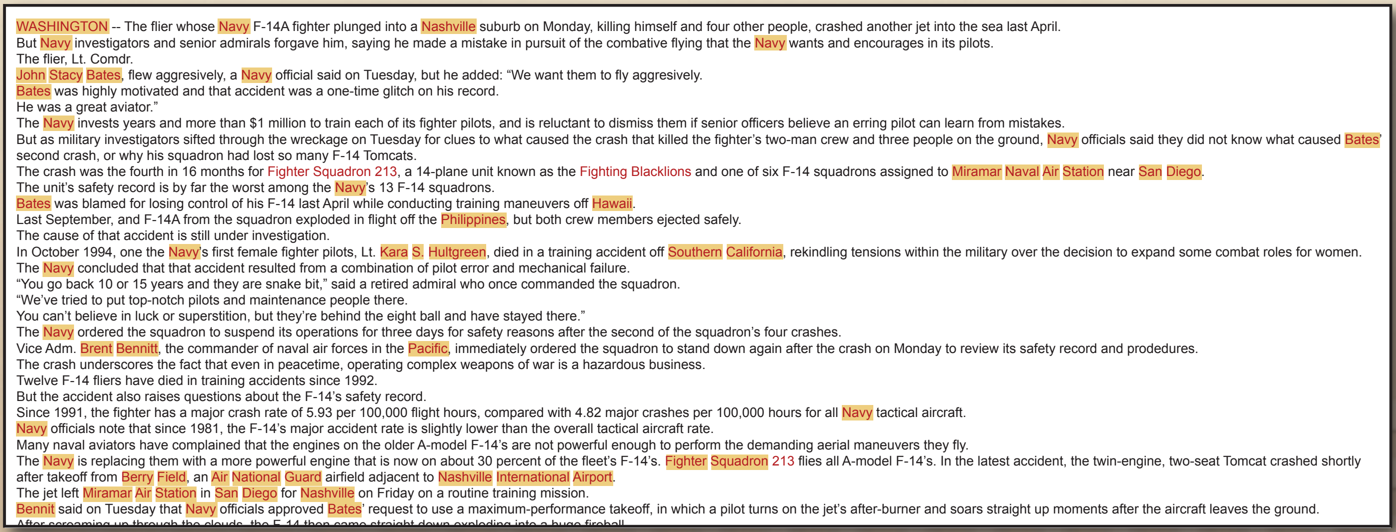
FY2009 Accomplishments and Results

The entity extractor aggregation tool has been constructed to serve as both a prototype of existing aggregation methodologies (meta-extractors) and an environment to enable future advancements. Four types of meta-extraction methodologies have been prototyped within the tool, each consisting of various interlinked modular components that include 1) extractor error detection; 2) error probability estimation; 3) hypothesis and likelihood computation; and 4) input/output of entity data. Each component relies on user-specified parameters that determine its behavior within a given algorithm (Fig. 1). For example, the user may specify the desired error space (the specific errors of interest) for the extractor error detection and probability estimation components. Execution of meta-extraction algorithms has been fully automated within the tool, and queuing of experiments has been enabled.

Performance estimation for the base and meta-extractors takes place via cross-validation, in which the data must be partitioned into multiple folds with associated performance estimates that are typically averaged to obtain an overall estimate. The aggregation tool runs these folds in parallel for increased efficiency. When an algorithm has been completed, the user is provided with an array of statistics associated with the execution. These include 1) error counts and probability estimates for the base and meta-extractors; 2) detailed output of the entities extracted by the base extractors, the space of hypothesized ground truths proposed by the meta-extractor, and the corresponding meta-extractor result; 3) the rate that events of interest occur (*e.g.*, the frequency with which the meta-extractor recreates the truth when all base extractors fail); and 4) the original text with extracted and ground truth entities highlighted (Fig. 2). This information collectively provides substantial insight into the behaviors and performance of meta-extraction methodologies, enabling the potential for algorithm optimization and enhancement.

Related References

1. Chen, M., Q. Shao, and J. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer, 2000.
2. Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, **2**, 12, pp. 1137–1143, 1995.



WASHINGTON -- The flier whose Navy F-14A fighter plunged into a Nashville suburb on Monday, killing himself and four other people, crashed another jet into the sea last April.
But Navy investigators and senior admirals forgave him, saying he made a mistake in pursuit of the combative flying that the Navy wants and encourages in its pilots.
The flier, Lt. Comdr.
John Stacy Bates, flew aggresively, a Navy official said on Tuesday, but he added: "We want them to fly aggresively.
Bates was highly motivated and that accident was a one-time glitch on his record.
He was a great aviator."
The Navy invests years and more than $1 million to train each of its fighter pilots, and is reluctant to dismiss them if senior officers believe an erring pilot can learn from mistakes.
But as military investigators sifted through the wreckage on Tuesday for clues to what caused the crash that killed the fighter's two-man crew and three people on the ground, Navy officials said they did not know what caused Bates' second crash, or why his squadron had lost so many F-14 Tomcats.
The crash was the fourth in 16 months for Fighter Squadron 213, a 14-plane unit known as the Fighting Blacklions and one of six F-14 squadrons assigned to Miramar Naval Air Station near San Diego.
The unit's safety record is by far the worst among the Navy's 13 F-14 squadrons.
Bates was blamed for losing control of his F-14 last April while conducting training maneuvers off Hawaii.
Last September, and F-14A from the squadron exploded in flight off the Philippines, but both crew members ejected safely.
The cause of that accident is still under investigation.
In October 1994, one the Navy's first female fighter pilots, Lt. Kara S. Hultgreen, died in a training accident off Southern California, rekindling tensions within the military over the decision to expand some combat roles for women.
The Navy concluded that that accident resulted from a combination of pilot error and mechanical failure.
"You go back 10 or 15 years and they are snake bit," said a retired admiral who once commanded the squadron.
"We've tried to put top-notch pilots and maintenance people there.
You can't believe in luck or superstition, but they're behind the eight ball and have stayed there."
The Navy ordered the squadron to suspend its operations for three days for safety reasons after the second of the squadron's four crashes.
Vice Adm. Brent Bennitt, the commander of naval air forces in the Pacific, immediately ordered the squadron to stand down again after the crash on Monday to review its safety record and prodedures.
The crash underscores the fact that even in peacetime, operating complex weapons of war is a hazardous business.
Twelve F-14 fliers have died in training accidents since 1992.
But the accident also raises questions about the F-14's safety record.
Since 1991, the fighter has a major crash rate of 5.93 per 100,000 flight hours, compared with 4.82 major crashes per 100,000 hours for all Navy tactical aircraft.
Navy officials note that since 1981, the F-14's major accident rate is slightly lower than the overall tactical aircraft rate.
Many naval aviators have complained that the engines on the older A-model F-14's are not powerful enough to perform the demanding aerial maneuvers they fly.
The Navy is replacing them with a more powerful engine that is now on about 30 percent of the fleet's F-14's. Fighter Squadron 213 flies all A-model F-14's. In the latest accident, the twin-engine, two-seat Tomcat crashed shortly after takeoff from Berry Field, an Air National Guard airfield adjacent to Nashville International Airport.
The jet left Miramar Air Station in San Diego for Nashville on Friday on a routine training mission.
Bennit said on Tuesday that Navy officials approved Bates' request to use a maximum-performance takeoff, in which a pilot turns on the jet's after-burner and soars straight up moments after the aircraft leaves the ground.
After screaming up through the clouds, the F-14 then came straight down exploding into a huge fireball.

Figure 2. Highlighted extracted and ground truth entities.